Difference of Proportions

- **Categorical rather than interval data**
  - Nominal or ordinal measurements
  - So, you can't do math (inc. can't subtract) w/ the values
  - So, a proportion rather than a mean
- **Comparing values for 2 groups (subsamples)**
  - So… very much like the t-test from Lab 17
  - But not comparing *means*
- **Still asking whether the difference between the 2 groups is statistically significant**
  - So… just like every other statistical test
  - And, like each test, what's new is the standard error – how much the test statistic varies across samples

SOC424 – Statistics w/ Dr. Ellis Godard

Differences in Proportions

Example Comparison

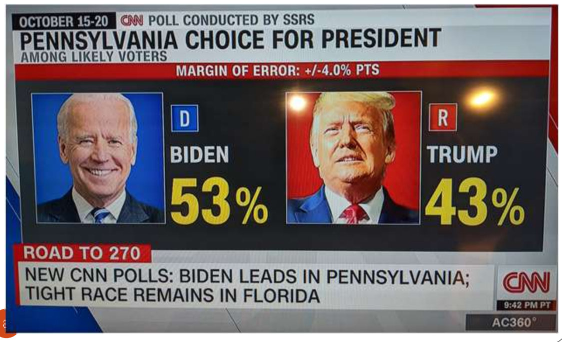
- **Some contend that the invasion of Iraq was supported due to "false beliefs"**
  - About 70% of Americans support war on Iraq
  - But 68% believe Iraq provided support to Al Qaeda
- **Research Question: Was support for the war higher among those with that belief?**
  - Of those who believe it, 74% support the war
  - Of those who don't, only 45% support the war
- **Statistical question: Was the difference between 74% and 45% statistically significant?**
  - Short answer: Yes, by a long shot

Outline for Today…

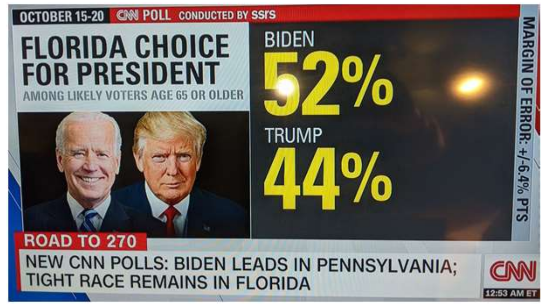
- **Difference of Proportions**
  - Hypotheses
  - Sampling Distribution
  - Standard Error
  - Test Statistic
- **Lab**

11/8/22 11pm

CIs don't overlap in PA (10/21/20)



No sig diff among older voters



SOC424 @ CSUN w/ Dr. Ellis Godard

11

CIs overlap in FL (10/21/20)

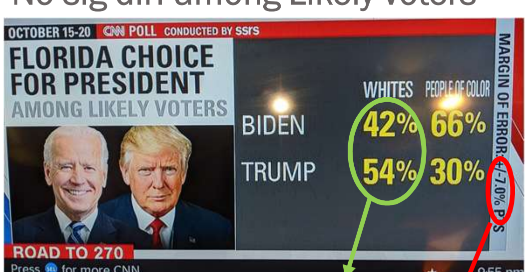
9

No Sig diff among Indies

SOC424 @ CSUN w/ Dr. Ellis Godard

12

No sig diff among Likely voters



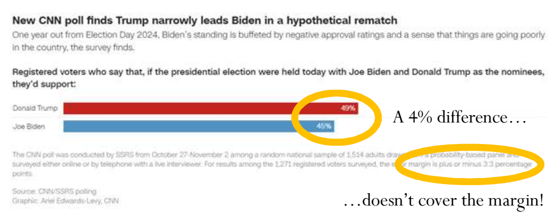
Observed 12% diff

But *each* could be off 7% (14 total)

SOC424 @ CSUN w/ Dr. Ellis Godard

10

Is Trump really leading Biden?



A 4% difference…

…doesn't cover the margin!

11/7/23

49 +/- 3.3 > 45.7 to 52.3

45 +/- 3.3 > 41.7 to 48.3

Those intervals OVERLAP!

SOC424 w/ Dr. Ellis Godard -- Slide

13

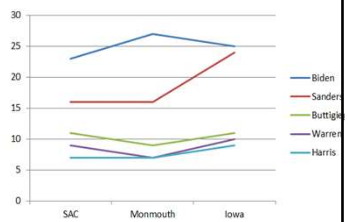
CA) are just a couple points behind him in all of these polls. But it's worth noting when a candidate suddenly goes from regularly polling at zero to 1 percent to placing third in a very large field.

Those three recent polls are:

- **Saint Anselm College released a poll** of the New Hampshire primary showing Biden with 23 percent, Sanders 16 percent, Buttigieg 11 percent, Elizabeth Warren 9 percent, and Kamala Harris at 7 percent.
- **Monmouth released a poll of Iowa** showing Biden in the lead with 27 percent, Sanders at 16 percent, Buttigieg with 9 percent, and Warren and Harris with 7 percent each.
- The **most recent Iowa poll before that, from Emerson** last month, showed Biden with 25 percent, Sanders 24 percent, Buttigieg 11 percent, Harris 10 percent, and Warren 9 percent.

Whether Buttigieg will manage to retain this position or grow his support further is of course unclear at this point, but we now have three polls suggesting he's struck a chord with some Democratic voters in Iowa and New Hampshire in a way that most of the many other Democratic candidates who have jumped into

	SAC	Monmoutl	Iowa
Biden	23	27	25
Sanders	16	16	24
Buttigieg	11	9	11
Warren	9	7	10
Harris	7	7	9



Setting up a Hypothesis Test

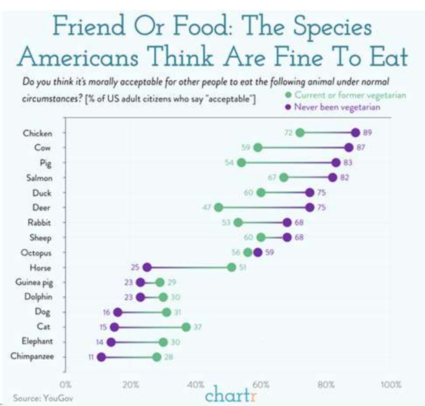
Usually, we are interested in whether or not the difference between two proportions is equal to zero. That is, for a two-tail test,

$H_o: \pi_1 - \pi_2 = 0$ or $H_o: \pi_1 = \pi_2$

$H_a: \pi_1 - \pi_2 \neq 0$ $H_a: \pi_1 \neq \pi_2$

These are Equivalent Statements.

17 SOC424 @ CSUN - Ellis Godard



Chartr, 9/3/23

15

Variance of the Diff. Of Prop.

- **Technical wording:**
  - If two estimates are formed from two independent samples, the variance of the sampling distribution for their difference (or their sum) is the sum of the variances of the sampling distributions of the two separate estimates (Agresti and Finlay, p. 168).
- **Practical translation**
  - You can't use that "p times one minus p" thing; you gotta use a more confusing one, ugh! (Godard, slide 18)

18 SOC424 @ CSUN - Ellis Godard

Standard Error of the Difference of Proportions

Standard deviation, is the square root of the variance (since the variance is the standard deviation squared)

$$\sigma_{\pi_1\text{-}\pi_1} = \sqrt{\sigma^2_{\pi_1\text{-}\pi_1}}$$

$$= \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}$$

Information from 1st group Information from 2nd group

19 SOC424 @ CSUN - Ellis Godard

To Actually Compute the Standard Error with Available Information

Replace π with p to estimate standard error

$$\hat{\sigma}_{\pi_1\text{-}\pi_1} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

20 SOC424 @ CSUN - Ellis Godard

Test Statistic

Since the hypothesized difference is zero, that part falls out – and the formula simplifies to something that looks like z again, yay ☺

$$\text{"z" (really t)} = \frac{\text{Estimate - Null Hypothesis Value}}{\text{Standard Error}}$$

$$= \frac{p_1 - p_2 - 0}{\hat{\sigma}_{p_1 - p_2}} = \frac{p_1 - p_2}{\hat{\sigma}_{p_1 - p_2}}$$

Once you have (& interpret) the test statistic, find the p value, interpret it, & state your conclusion.

21 SOC424 @ CSUN - Ellis Godard

Hypotheses (Difference of Proportion)

H_0: $\pi_S = \pi_N$ (Dropout Rate is the Same in South and the rest)

H_a: $\pi_S > \pi_N$ (Dropout Rate Higher in South than in the rest)

Output from GSS
Regional Differences in Dropout Rates

Region	Proportion	Std Error
South (n=411)	0.292	0.022
Non-South (n=1099)	0.197	0.012

22 SOC424 @ CSUN - Ellis Godard

Computing the Test Statistic

First, find the standard error …

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

$$= \sqrt{\frac{0.292(1-0.292)}{411} + \frac{0.197(1-0.197)}{1099}}$$

$$= 0.0254$$

23 SOC424 @ CSUN - Ellis Godard

Computing the Test Statistic

Then compute the test statistic itself:

$$z = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} = \frac{0.292 - 0.197}{0.0254} = 3.74$$

The corresponding *p-value,* for a one tail test, is between 0.000233 and 0.0000317 (Table A). What do we conclude?

24 SOC424 @ CSUN - Ellis Godard

Example from SPSS

- **Reminder: SPSS treats all data as interval**
  - Calculates a mean for a nominal variable
    - Also calculates S.E. and performs t-tests, *as if* the data were interval
  - *None of that* is meaningful for categoricals, which *don't have means!*

  Therefore, you must get the *pieces* you need from SPSS,
  then calculate by hand or w/ a calculator (and *show your work!!*)
- **So, what pieces do you need?**
  - The two *valid* proportions, and the 2 *valid* frequencies – that's it ☺
  - NOT the mean. NOT the standard deviation. NOT the S.E. Mean.

25 SOC424 @ CSUN - Ellis Godard

Example from SPSS, contd

- **Do more students do art or athletics?**
  - survey.sav has q28_art & q28_athl
  - 16.7% (.167) are involved in athletics
  - 4.2% (.042) are in art, music or drama
  - Is the difference between those proportions statistically significant?

Example from SPSS, contd

$$\hat{\sigma}_{\pi_{athl} - \pi_{art}} = \sqrt{\frac{.042\,(1-.042)}{24} + \frac{.167\,(1-.167)}{24}}$$
$$= \sqrt{\frac{.040236}{24} + \frac{.139111}{24}}$$
$$= \sqrt{.0016765 + .0057963}$$
$$= \sqrt{.00745613} = .08635$$

Example from SPSS, contd

$$z = \frac{\hat{\pi}_{athl} - \hat{\pi}_{art}}{\hat{\sigma}_{\pi_{athl} - \pi_{art}}} \approx \frac{.042 - .167}{.08535}$$
$$= \frac{-.125}{.08535} = -.14646$$

The observed proportions differ by an estimated .14646 standard errors. The p-value associated with that z statistic is far greater than .05 (see Table A).

The risk of being wrong in rejecting the null is too high, we are unable to reject the null, and we are unable to conclude that the proportions are significantly different

Proportional Diffs by Gender

DEMO

- **Conduct a hypothesis test using GSS**
  - Question: Are men or women more likely to be widowed?
  - Use gss84.sav, & get two sets of *univariate* data – one per group
- **Select male cases**
  - Data / Select Cases / If Condition Satisfied / Sex =1
  - Get proportion that are widowed
- **Select female cases**
  - D / SC / ICS / Sex = 2 (all cases first?)
  - Get proportion who are widowed
- **Conduct test (assumptions, hypos, stat, p value, & conc.)**

Example from SPSS, contd

$$\hat{\sigma}_{\pi_{Men} - \pi_{Female}} = \sqrt{\frac{.033(1-.033)}{598} + \frac{.153(1-.153)}{875}}$$
$$= \sqrt{\frac{.031911}{598} + \frac{.129591}{875}}$$
$$= \sqrt{.000053622 + .000148104}$$
$$= \sqrt{.000201726} = .01420$$

$$t = \frac{\hat{\pi}_{men} - \hat{\pi}_{women}}{\hat{\sigma}_{\pi_{men} - \pi_{women}}} \approx \frac{.033 - .153}{.01420}$$
$$= \frac{-.120}{.01420} = -8.45428$$

For Your Next Lab…

- **Conduct a hypothesis test using GSS**
  - Question: Are males more likely to get 12 or more years of education?
  - Use two variables in gss84.sav: SEX and EDUC (*not* degree)
- **Select the male cases**
  - Data / Select Cases / If Condition Satisfied / Sex =1
  - Get proportion with *12 or more* yrs Educ
    - Hint: *start with* the cum % for 11
  - Do *not* use a mean or standard deviation
    - This reading, lecture, notes, quiz, & lab are about *proportions*
- **Select female cases**
  - D / SC / "all cases" first, to undo the above filtering
  - D / SC / ICS / Sex = 2
  - Get proportion with 12 or more yrs Educ
- **Conduct test - all 5 steps:**
  - Assumptions, hypotheses, test statistic*, p value*, and conclusions
    * including interpretations